Received: 23 April 2023 | Revised: 19 June 2023 | Accepted: 3 July 2023
DOI: 10.1002/psp4.13014

ARTICLE

Can a propensity score matching method be applied to assessing efficacy from single-arm proof-of-concept trials in oncology?

Shu-Wen Teng[1] | Yu-Cheng Su[1] | Ravikumar Pallantla[2] | Madhav Channavazzala[2] | Rukmini Kumar[2] | Yucheng Sheng[1] | Hao Wang[1] | Crystal Wang[1] | Archie Tse[1]

[1]CStone Pharmaceuticals, Su Zhou, China
[2]Vantage Research, Inc., Lewes, Delaware, USA

Correspondence
Shu-Wen Teng; Department of Research, CStone Pharmaceuticals, Su Zhou, China.
Email: shuwenteng@gmail.com

Abstract

As a result of the escalating number of new cancer treatments being developed and competition among pharmaceutical companies, decisions regarding how to proceed with phase III trials are frequently based on findings from either single-arm phase I expansion cohorts or phase II studies that compare the efficacy of the study drug to a standard-of-care benchmark derived from historical data. However, even when eligibility criteria are matched, differences in the distribution of baseline patient features may influence the outcome of single-arm trials in real-world scenarios. Therefore, novel methods are needed to enhance the accuracy of efficacy prediction from current cohorts relative to historical data. In this study, we demonstrated the feasibility of using the propensity score matching (PSM) method to improve decision making by matching relevant baseline features between current and historical cohorts. According to our findings, utilizing the PSM method may provide a less biased means of comparing outcomes between current and historical cohorts relative to a naïve approach, which relies solely on differences in average outcomes between the cohorts.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

As the number of new cancer treatments entering clinical trials continues to rise, accompanied by increased competition among pharmaceutical companies, decisions regarding phase III trials are sometimes based on findings from single-arm proof-of-concept (PoC) trials alone. These decisions rely on naive comparisons, which are estimated through differences in average outcomes between a cohort in a single-arm trial for the novel treatment and those in historical cohorts.

WHAT QUESTION DID THIS STUDY ADDRESS?

To address the impact of differing baseline patient characteristic on outcomes in real-world scenarios, this study sought to explore the suitability of using a propensity score matching (PSM) method in evaluating efficacy in single-arm PoC trials.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

We successfully demonstrated the feasibility of using PSM to match relevant baseline features between current and historical cohorts. Our findings indicate that PSM provides a more objective means for comparing outcomes between cohorts in contrast to naive comparisons.

HOW MIGHT THIS CHANGE DRUG DISCOVERY, DEVELOPMENT, AND/OR THERAPEUTICS?

PSM may offer a more informative approach to evaluating the efficacy of a new therapy from single-arm PoC trials as it addresses potential confounding factors arising from imbalanced predictive baseline features.

INTRODUCTION

In oncology drug development, proof-of-concept (PoC) data on drug safety and efficacy need to be established in early-phase clinical trials before large phase III randomized controlled trials (RCTs) are conducted. PoC trials (among other variable design considerations) can be broadly conducted as single-arm trials or RCTs, with clear advantages and disadvantages to either approach. For every RCT, approximately four single-arm trials can be run instead, limiting the number of PoC hypotheses to be tested.[1] With the increasing number of novel cancer therapies in development and competition among pharmaceutical companies, especially in the immune-oncology space, a "go" decision to phase III is increasingly made based on results from a single-arm phase I expansion cohort or phase II study in which efficacy of the study drug (e.g., objective response rate [ORR]) is compared with standard-of-care benchmark estimated from historical data.[2] This practice has unfortunately led to several costly phase III failures.[3–5]

In cases where a cohort in a single-arm trial for the novel treatment and historical cohorts have similar patient distributions, a naïve comparison estimated through differences in average outcomes of the cohorts, may suffice to make the decision. However, in most real-world scenarios, despite having matching eligibility criteria, the risk of having a different distribution of known and unknown baseline patient features that may influence trial outcome remains significant. New approaches[6] are required to improve the precision of outcome comparison between current and historical cohorts, as shown in Figure 1.

An emerging set of statistical and machine learning approaches are being used for evaluating the comparative effectiveness of an intervention using subsets of patients derived from external data.[7] The source of these external control data can be from other clinical trials, electronic health records, or published aggregated data.[8] Statistical methods include propensity score matching (PSM),[9] fixed-effect pooling, multivariate regression, and Bayesian dynamic borrowing. These statistical methods focus on selecting a subgroup of patients from the external data, which minimizes the imbalance of patient characteristics between the external control arm and the current trial cohort. Other methods, such as machine learning methods, use predictive or generative models to artificially generate patient response data and then create the external control arm.[10] In late-phase clinical development, externally controlled trials have been used globally to support regulatory approval[11] of new therapies for rare diseases and oncology.[12]

Within the context of early oncology drug development, here, we demonstrated the superiority of the PSM method over a naïve approach and quantified the extent of this superiority in facilitating a go/no-go decision based on single-arm PoC trials. Specifically, we evaluated the added value of the PSM method in two scenarios: intra-trial and inter-trial comparison of current and historical cohorts. Our findings demonstrate that the PSM method provides a more reliable comparison of the outcome between a current and historical cohort in contrast to a naïve comparison, which solely relies on differences in average outcomes of the cohorts.

METHODS

General approach

We established the utility of the PSM method with two case studies using clinical data from the non-small cell lung cancer (NSCLC) population. In case study 1, we relied on data derived from the experimental arm of GEMSTONE-302 (NCT03789604), a randomized phase III trial comparing the efficacy of the combination of a novel PD-L1 monoclonal antibody sugemalimab (CS1001)

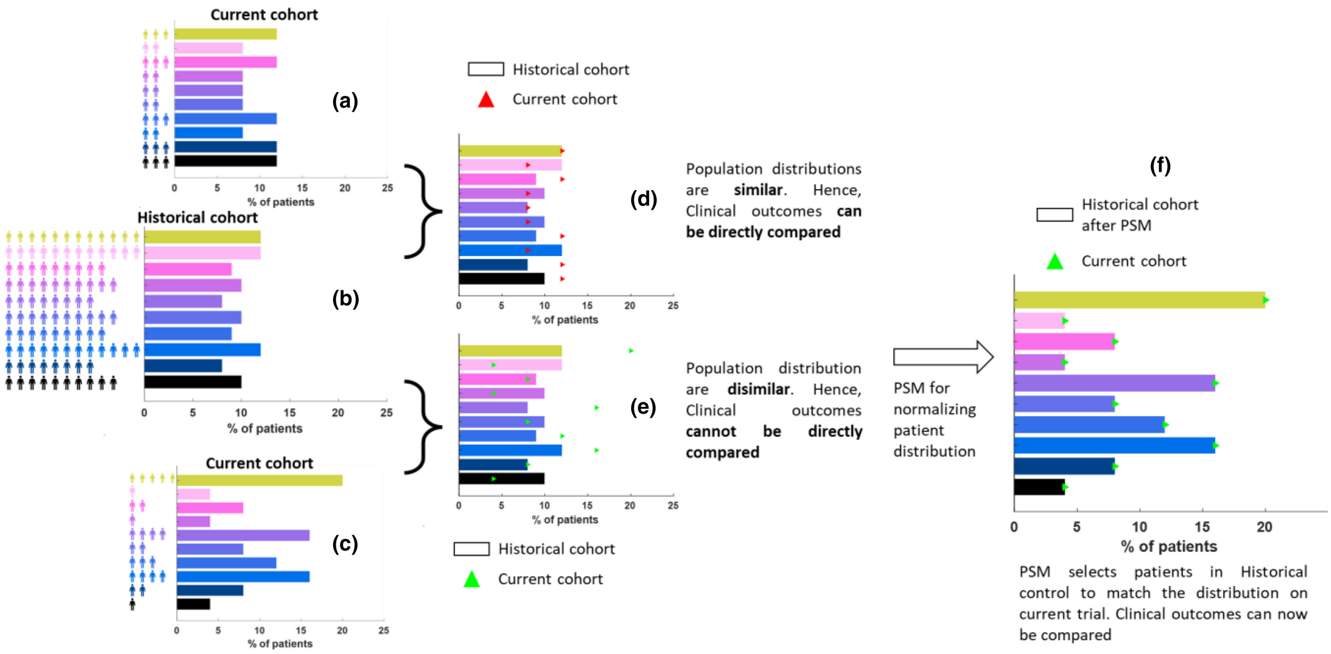

FIGURE 1 Illustration of two scenarios encountered when comparing baseline features distribution between current and historical cohorts. (a) Current cohort population (similar to the historical cohort). (b) Historical cohort population. (c) Current cohort population (dissimilar to the historical cohort). (d,e) Overlaid historical and current cohort distributions. (f) Propensity score matching (PSM) method was used to match baseline features between current and historical cohorts. Colors represent the distinction in their hypothetical baseline features (demographics/biomarkers).

plus chemotherapy with placebo plus chemotherapy for patients with metastatic NSCLC.[13] Patients were split into two cohorts with a predefined difference in outcome distribution: "current" and "historical" cohorts (an intra-trial comparison). The clinical outcomes were ORR and progression-free survival (PFS), as defined per Response Evaluation Criteria in Solid Tumors (RECIST) version 1.1.[14] In case study 2, we applied the PSM method for an inter-trial comparison using a phase I trial as the current cohort and the above-mentioned phase III trial as the historical cohort (an inter-trial comparison). In both case studies, we used PSM[9] to create external control arms. PSM is a practical statistical approach that is suited for dealing with the small patient sizes commonly encountered in early phase clinical trials.[15,16] We evaluated the difference between the magnitude of treatment effect derived from the PSM method and that obtained from a naive approach. The common procedure used for the PSM analysis in both case studies is described in the flow chart (Figure S1).

Dataset and the generation of "historical" and "current" cohorts

Case study 1 used data from the GEMSTONE-302 trial, which had 320 patients receiving sugemalimab plus chemotherapy and 159 patients receiving placebo plus chemotherapy. Only patients in the sugemalimab plus chemotherapy group were analyzed, resulting in 303 patients after excluding 17 with missing ORR information (Figure S1). Each patient had 64 baseline features, including demographic and biomarker information at baseline. Utilizing only baseline features is crucial to avoid the potential introduction of selection biases when matching on post-baseline features. Table S1 summarizes the features and their descriptions. The historical and current cohorts were created by splitting the sugemalimab plus chemotherapy group, with a predetermined variation in outcome distribution (Figure 2a). This splitting was solely for the purpose of simulating a partial mismatch between historical and current data and not necessary for comparing data from different trials in real-world scenarios.

The patients were split based on an estimated parameter of tumor growth kinetics – the log of on-treatment tumor growth parameter, LKG.[17,18] LKG was selected because of its mechanistically explainable influence on the patient clinical outcomes; and in fact, in this cohort, LKG correlated well with the clinical outcomes, including ORR, PFS, and overall survival (unpublished data). Further details of this splitting are given in Creation of historical and current cohorts. Because the main objective of this case study is to evaluate the PSM method, any parameter that is influenced by patients' baseline features has the potential to be utilized for the splitting, as illustrated in the Discussion.

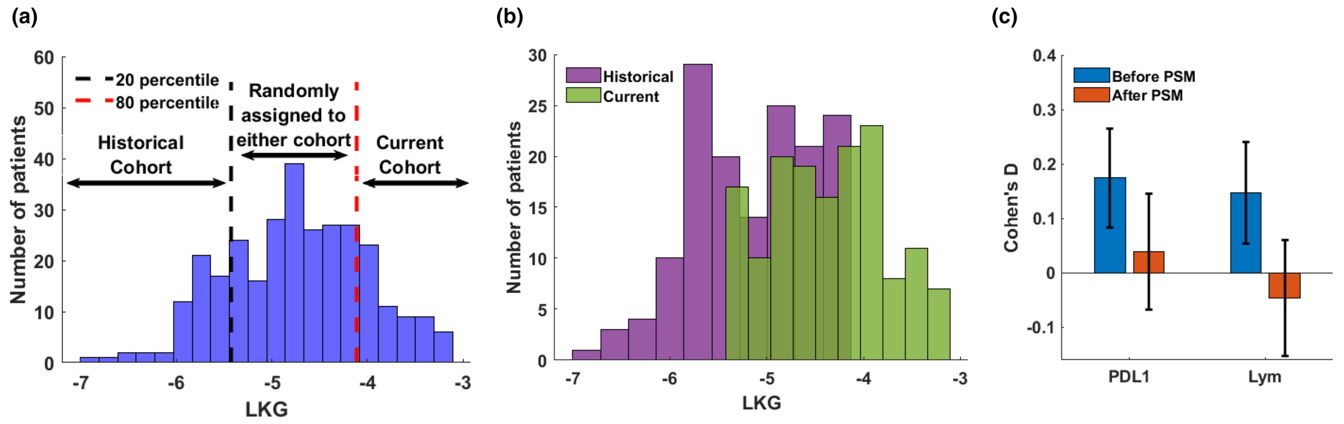

FIGURE 2 Data splitting approach for case study 1 (intra-trial comparison). (a) Distribution of logarithm of tumor growth (LKG) of sugemalimab group of GEMSTONE-302. (b) Distribution of LKG for the current and historical cohorts created through our splitting method. (c) Cohen's D of the key features (PDL1 and Lym) before and after propensity score matching (PSM). The bar shows the mean value and error bars show the standard deviation from 1000 iterations.

In case study 2, we used data from GEMSTONE-302 as the historical cohort and data from a phase I trial NCT03312482[19] as the current cohort. NCT03312482 is the first-in-human phase I trial of sugemalimab monotherapy or in combination with chemotherapy in patients with various advanced malignancies. For our analyses, we selected patients who received sugemalimab plus chemotherapy from the NSCLC cohort, an indication matching that of GEMSTONE-302. This cohort contained 41 patients, out of which 24 patients had missing PD-L1 biomarker information. Hence, the remaining 17 patients were used as the "current" cohort, as shown in Figure S1.

Patient feature selection and propensity score

To begin with, we selected the subset of features from the available 64 features. Our selection was based on their relevance to the treatment outcome ORR, taking into account both statistical relationships and biological significance, as indicated by refs. 9 and 20. The features were linearly transformed to have a mean of zero and standard deviation of unity. We used the logistic regression with least absolute shrinkage and selection operator (LASSO) regression method to understand the feature importance (known from weights) in predicting ORR. We chose LASSO regression as it provides good out-of-sample prediction even in the presence of multicollinearity between features. In Identification of patient features relevant to clinical outcome, we discussed other methods that were also considered. For the feature selection process, we utilized data from the sugemalimab plus chemotherapy arm in the GEMSTONE-302 dataset, which is commonly used in both case studies. We selected features with high absolute weights and strong correlation to ORR to develop a metric that characterizes the patients (Figure S2).

Propensity score is the probability that a patient would belong to a particular group given a set of baseline features.[21] It indicates how strongly the baseline features influence the patient belonging to the current treatment group. If the baseline features are highly influential, the current and historical cohorts can be easily separated based on their features. If there is a clear differentiation between two cohorts based on baseline features, it is possible to distinguish the distribution of outcomes. This distinction can occur if the baseline features possess the capability to predict outcomes and no other confounding factors exert influence on the outcome. Figure S3A-D illustrate the process for calculating propensity scores in case study 1, and the logit of propensity score is used for patient matching.[22]

Criterion and performance metric of patient matching

Propensity scores were used for patient matching between the current and historical cohorts. For every patient p_i in the current cohort, we identified a patient q_j from the historical cohort with the closest propensity score. We matched q_j with p_i if the absolute difference between their propensity scores was less than a predetermined caliper width.[23] The literature suggests a caliper width of 20% of the standard deviation of all patients' scores.[24] However, this was too weak for our dataset as it resulted in matching patients with very different baseline characteristics. We gradually reduced the it and found that a 1% standard deviation was appropriate, as a further reduction led to a significant reduction in matched patients. Therefore, we used a 1% standard deviation as our caliper width for patient matching.

Regarding the matching process, two observations were made: (1) if a patient in the current cohort did not find an appropriate match with any patient in the historical cohort, this patient was dropped from the current cohort. This observation was rare and did not significantly impact current cohort sizes. It is worth noting that whereas PSM is a commonly used method, alternative approaches like inverse propensity treatment weighting incorporate all patients without excluding any potential unmatched patients.[25] (2) It was possible for a single patient from the historical cohort to be matched with multiple patients in the current cohort.[15,16]

The difference in the distributions of a baseline feature between the historical and current cohorts is defined using the dimensionless Cohen's D,[26] a standardized mean difference metric. The mathematical description of Cohen's D is in Supplementary Section S2 and a pictorial representation is shown in Figure S3E,F. A small absolute value of Cohen's D suggests that the distributions have a large overlap, whereas a large absolute value of Cohen's D suggests that the distributions of baseline features are separated and have little or no overlap. The better the performance of patient matching, the smaller the absolute value of Cohen's D between "historical cohort after PSM" and "current cohort before PSM."

RESULTS

Creation of historical and current cohorts

In case study 1, the historical and current cohorts were created from LKG-based splitting of the data derived from the sugemalimab-treated arm of GEMSTONE 302, containing 151 and 152 patients, respectively. In this splitting, patients in the lower 20 and upper 20 percentile of LKG were respectively assigned to the historical and current cohorts. The remaining patients were randomly assigned to the two cohorts in equal ratios. The purpose of doing splitting was to establish a distinction in outcome (i.e., ORR) among patients participating in the same trial, in which patients who responded better were intentionally allocated to one cohort instead of the other. By randomly assigning the intermediary 60% of patients, 1000 iterations were generated. The distribution of LKG in the two cohorts before and after splitting is shown in Figure 2a,b. We also analyzed the sensitivity of the PSM output to this splitting criterion, which is explained in Effect of overlap between cohorts on PSM performance.

Identification of patient features relevant to clinical outcome

In identifying the outcome-relevant subset of baseline patient features from the original set of 64, we observed that both forward and backward sequential feature selection methods (described in Supplementary Section S1) were unstable in our case, that is, the selected features are completely different when one of the patients is randomly dropped from the dataset (as shown in Table S1). Hence, these methods are considered inappropriate for feature selection in our case. In contrast, LASSO regression was found to be stable, as evident from Table S1 that compares the features and their weights between those cases when a patient is randomly dropped/not dropped from the dataset. In addition to the importance assigned to the feature by the LASSO method, we also evaluated the correlation between the feature and the clinical outcome. This was to guarantee that our feature selection is not entirely dependent on the selection method employed. Figure S2 shows the relation between the feature weights estimated from LASSO and their respective correlation coefficients to the clinical outcome. The relevance of a feature as dictated by LASSO seems higher if that feature's correlation with the outcome is also higher. Hence, we selected these set of features, namely: PDL1 (programmed death ligand-1 > 1%), Lym (lymphocyte count), NEOIWTYP (squamous cell carcinoma), K (potassium measurement), and PROT (total protein measurement). Among these, it is observed that using only two of the features (PDL1 and Lym) resulted in an optimal performance of PSM as explained in Effect of feature-outcome correlation on PSM performance. This is because the weakly correlated features (NEOIWTYP, K, and PROT) reduce the contribution of the truly important features (PDL1 and Lym) to the propensity score. More details on this and the effect of auto-correlations between the features on performance of PSM are discussed in Effect of feature-outcome correlation on PSM performance. Our analysis does not rule out the possibility that there are other unmeasured baseline features (confounding variables) that are relevant to the outcome.

Propensity scores were calculated from the patient cohorts generated using the relevant features identified as described in Section 2.3. Patients in the current cohort were matched with the historical cohort using the propensity scores, as shown in Figure S3A–D.

We used Equation S1 to compute the Cohen's D value for evaluating the discrepancy in distribution between the current and historical cohorts regarding two key features, PDL1 and Lym. To avoid the dependence of the Cohen's D value on a single random assignment of the splitting, we performed 1000 random assignments and determined the average and standard deviation of the resulting Cohen's D values (Figure 2c). The comparison of the Cohen's D values in Figure 2c indicates a decrease after PSM, signifying the effectiveness of PSM in reducing the baseline feature differences between the historical and current cohorts.

Results from case study 1 (intra-trial comparison)

For case study 1, the disparity before PSM between historical and current cohorts in the clinical outcomes ORR and PFS can be seen in Figure 3a,b, respectively. This disparity was introduced by our splitting strategy. We hypothesized that PSM could reduce this disparity and therefore allow a more informative outcome comparison between the cohorts. Our findings provide evidence that PSM reduced, but did not eliminate, the disparity in outcomes between the historical and current cohorts. Thus, the ORR values in the historical cohort before and after PSM are 74% and 64%, respectively; the ORR values in the current cohort before and after PSM are 54% and 53%, respectively (Figure 3a). The median PFS values in the historical cohort before and after PSM are 7.1 and 6.4 months, respectively; the median PFS values in the current cohort before and after PSM are 5.6 and 5.7 months, respectively (Figure 3b). Note that the minimal difference found in the current cohort before and after PSM implies that most patients in the current cohort were suitably matched with those in the historical cohort. Therefore, the current cohort distribution remained stable while undergoing PSM treatment. The evaluation of PSM feasibility should use the outcome of the current cohort before PSM, as this represents the observed current cohort outcome that PSM aims to predict. The decrease in Cohen's D value for both ORR and PFS from "before PSM" to "after PSM" depicted in Figure 3c indicates that the use of the PSM method resulted in an improvement in the overlap of the outcome distributions between the current and historical cohorts. The findings from Figure 3c are consistent with those from Figure 3a,b.

Results from case study 2 (inter-trial comparison)

In case study 2 (inter-trial comparison), to recapitulate, we used phase I trial (NCT03312842) as the current cohort and phase III trial (sugemalimab group of GEMSTONE-302) as historical cohort as shown in Figure S1. The number of patients in these trials were different by an order of magnitude, with the sugemalimab group of GEMSTONE-302 containing 303 patients and NCT03312842 containing a mere 17 patients.

We were concerned that if we applied PSM to the full dataset of GEMSTONE-302 and NCT03312842, the linear classifier might be biased toward the former as it contained a greater number of patients. Instead, we randomly sampled 30 patients from the sugemalimab group of GEMSTONE-302 trial and used them as the historical cohort for PSM. An illustration of this process is provided in Figure S4A-D. The procedure was reiterated 10,000 times, and only the cases where a minimum of 75% of the patients in the current cohort achieving a match were retained. Approximately 350 applicable iterations emerged, and the findings presented are based on this subset. Figure S4E,F display a decrease in Cohen's D value for median PFS after the implementation of PSM, suggesting that the utilization of the PSM method led to a better alignment of the outcome distributions between the current and historical cohorts. As illustrated in Figure S4G, the difference in ORR between the historical and current cohorts before PSM appears to be insignificant (66% in the historical cohort and 67% in the current cohort) and this poses challenges in evaluating the feasibility of the PSM method for comparing ORR. After performing PSM, the

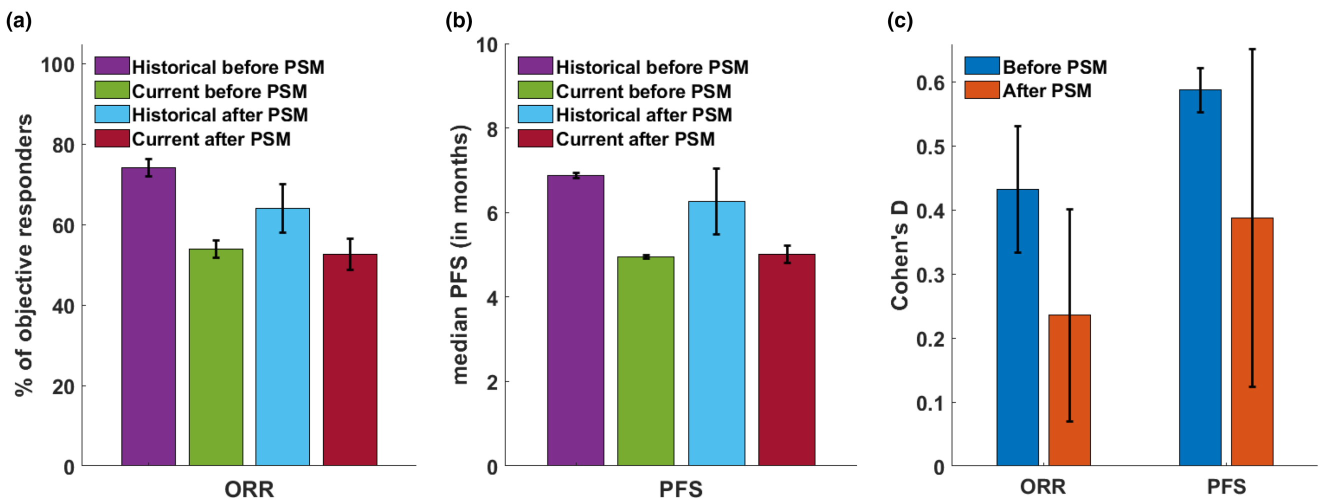

FIGURE 3 Results of propensity score matching (PSM) applied to sugemalimab group of GEMSTONE-302 trial (case study 1) using the features PDL1 and Lym. (a) Objective response rate (ORR) of current and historical cohorts before and after PSM treatment. (b) Median progression free survival (PFS) of current and historical cohorts before and after PSM treatment. (c) Cohen's D of ORR/PFS outcomes. The bar shows the mean value and error bars show the standard deviation from 1000 iterations.

ORR remains consistent (68% in the historical cohort and 69% in the current cohort). The preservation of ORR in the matched samples compared to the unmatched samples suggests that the PSM method does not introduce bias in scenarios where the current and historical cohorts have similar distributions (as shown in Figure 1a,b).

To summarize, the PSM method can be utilized to generate a comparison that minimizes the influence of confounding relevant baseline features, thereby reducing the disparities in the patient distribution between the historical and current cohorts.

DISCUSSION

In case study 1, we deliberately created two distinct distributions between the historical and current cohorts, with the anticipation that it would lead to a significant difference in outcomes when using a naïve comparison. The assessment of the PSM method's effectiveness would only be deemed feasible if the outcome difference decreases after applying PSM treatment. Our analyses showed that using the PSM method can lead to a more reliable comparison of the outcome between a current cohort and a historical cohort, as opposed to a naive comparison that relies on differences in average responses of the cohorts. Specifically, the utilization of the PSM method resulted in a reduction of the ORR disparity from 20% to 10% between the historical and current cohorts, as depicted in Figure 3a. Similarly, Figure 3b illustrates a decrease in the median PFS difference from 1.5 to 0.8 months between the historical and current cohorts with the application of the PSM method. The observed difference in ORR/PFS values between the historical cohort after PSM and the current cohort before PSM could be attributable to two possible factors: (1) relevant baseline features that impact the outcome may still exist but were not measured or included in the analysis (see Supplementary Section S4 and Figure S4H), or (2) the distribution of outcomes in the historical cohort does not fully represent the distribution of outcomes in the current cohort.

When an alternative splitting method is used, the conclusion that the PSM method provides a less biased comparison of outcomes remains unchanged. When ORR was used to split the data for the intra-trial comparison in case study 1, the same conclusion derived from the LKG-based splitting held true (Figure S5A-C). Both results (Figure 3 and Figure S5A-C) demonstrate the robustness of PSM method with respect to different data preparation approaches.

In case study 2, the historical cohort was defined as a phase III trial, whereas the current cohort was defined as a phase I study with the same indication. The expectation was that the distributions between these two cohorts could display similarities or differences. This variation is primarily due to the potential considerable variability resulting from the small sample size of the current cohorts ($N=17$), as described in the Supplementary Section S4 and illustrated in Figure S4H. If a significant difference in outcomes is observed between the historical and current cohorts, the assessment of the effectiveness of the PSM method would only be considered feasible if the outcome difference decreases after applying the PSM treatment. On the other hand, if there is no significant difference in outcomes, the feasibility cannot be assessed. In such cases, the assessment of the PSM method would only be considered non-harmful if the outcome difference remains unchanged after applying the PSM treatment. Our analyses revealed that the current and historical cohorts exhibit similar distributions. The application of the PSM method enables a less biased comparison of median PFS between the two cohorts (Figure S4E,F) without introducing bias in ORR (Figure S4G). Because there is already a good balance in the baseline characteristics between the two cohorts, this makes this case study supplementary in nature.

It is important to note that in case study 2, the phase III trial followed the phase I trial, unlike real-world scenarios where the current trial typically follows the historical one. Although hypothetical, this case study validates the PSM method for real-world scenarios involving cohorts with different sample sizes and paves the way for future applications of PSM.

In summary, we have demonstrated that the PSM method allows a less biased comparison of outcomes between the current and historical cohorts by normalizing the distributions of baseline features between the two cohorts, consequently reducing the confounding effects of patient variability. All the clinical trial data used in this study are based on the Chinese population. In other scenarios, regional differences in the patient population should be taken into consideration as it could play a significant role in observed treatment effect. The biological relevance of the two selected features, PDL1 and Lym, is also supported by the literature. For example, studies have established an association between absolute lymphocyte counts and therapy outcomes in various solid tumors.[27,28] In the case of patients with NSCLC, a higher baseline absolute lymphocyte counts showed better survival,[29] and peripheral lymphocyte count was identified as a surrogate marker of immune checkpoint inhibitor therapy outcomes.[30] Similarly, it is observed that immune checkpoint inhibitors showed better clinical outcomes with higher PD-L1 expression in patients with NSCLC.[31–34]

Our demonstration of the feasibility of the PSM method in case studies 1 and 2 highlights the need for further prospective validation studies to validate this approach. Whereas PSM holds greater recognition within the medical community, alternative approaches, such as counterfactual modeling, exist and used in practical settings.[35] To improve the performance of the PSM method and leverage its potential in facilitating decision making, we consider three data-related factors in the following discussion.

How to optimize the performance of PSM?

Effect of overlap between cohorts on PSM performance

PSM performance is optimized when there is a reasonable overlap of baseline features between historical and current cohorts, as depicted in Figure 4. Matching is challenging with little overlap between historical and current cohorts, yielding no improvement in the outcomes' Cohen's D values, as depicted in Figure 4a,d,g. When there is substantial overlap, PSM is not needed, and outcomes can be compared directly, as shown in Figure 4c,f,i. PSM is effective with significant differences but notable overlap in features, as seen in Figure 4b,e,h, which are based on the splitting method used in case study 1.

Effect of feature-outcome correlation on PSM performance

PSM does not guarantee similar treatment outcomes, as outcomes rely on the feature-outcome correlation. By

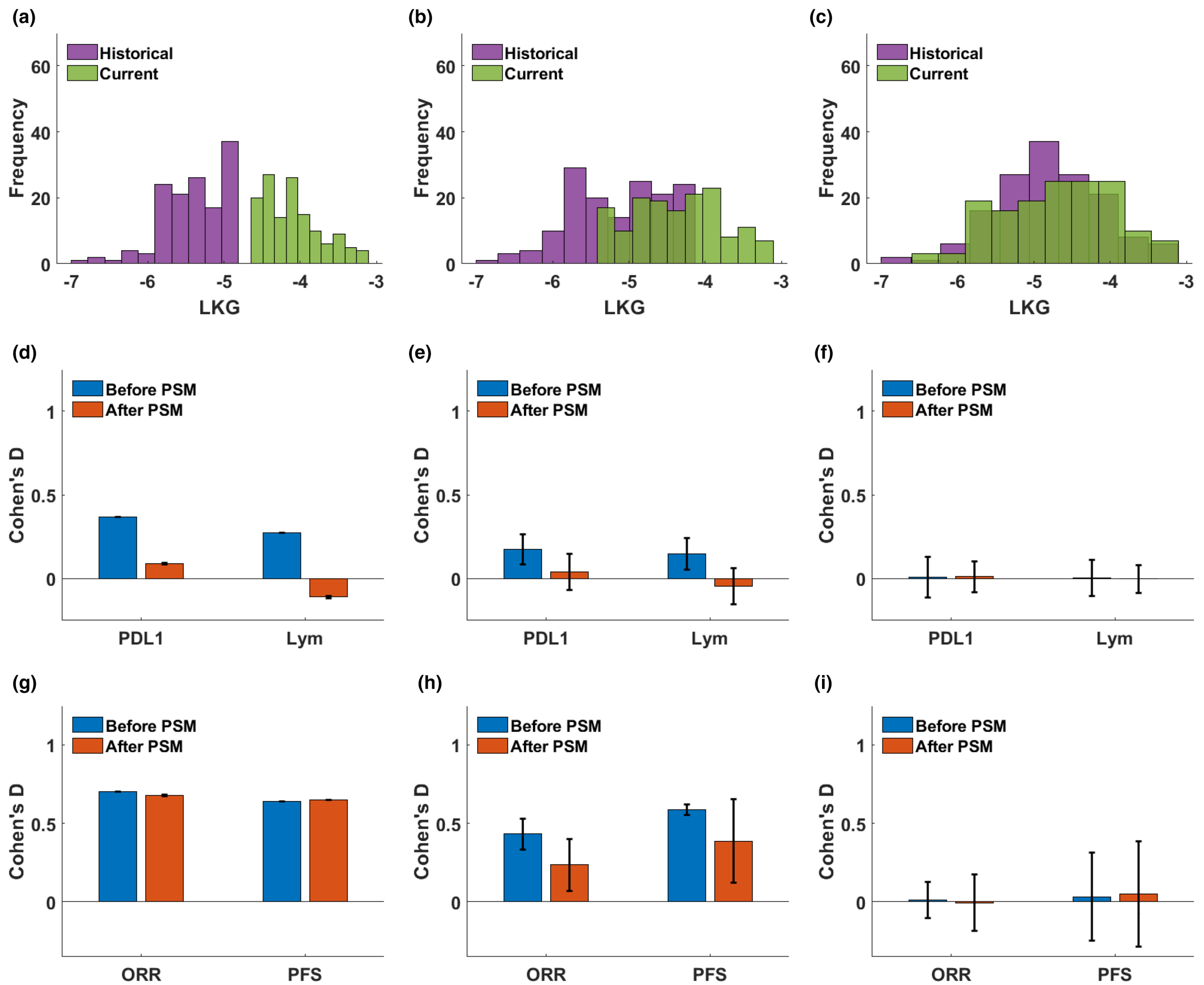

FIGURE 4 Different cases of splitting patients from sugemalimab group of GEMSTONE-302 in to current and historical cohorts (case study 1). (a-c) Representative histograms of logarithm of tumor growth (LKG) in the cases of no overlap, moderate overlap, and complete overlap respectively between the current and historical. (d-f) Cohen's D values for the two selected relevant features (PDL1 and Lym) before and after propensity score matching (PSM). (g-i) Cohen's D values for the two selected clinical outcomes, objective response rate (ORR), and progression-free survival (PFS), for the corresponding cohorts before and after PSM. Bars show the mean and error bars show the standard deviation from 1000 iterations.

using the five identified relevant features (Table 1), we compare the ORR in the historical and current cohorts before and after PSM (Figure S5D). The study shows that using only two highly correlated features, PDL1 and Lym, leads to better matching performance. However, including three weakly correlated features (NEOIWTYP, PROT, and K) lowers the performance.

To assess PSM performance with feature-outcome correlation and feature-feature correlation, we synthesized data due to insufficient real data for numerical experiments. Details on data synthesis are provided in Supplementary Section S3. In Figure 5a,b, the outcome of synthesized data is binary (analogous to ORR). In Figure 5a, there is only one continuous input feature (similar to Lym), whereas in Figure 5b, one input feature is continuous, and the other is binary (comparable to utilizing continuous Lym and binary PDL1 from actual data as input features). Figure 5a,b depict the impact of feature-outcome correlation on Cohen's D values. Increasing feature-outcome correlation leads to better predictability of PSM, resulting in a decrease in Cohen's D value. Cohen's D also depends on feature-feature correlation (Figure 5b), with a modest decrease observed when two input features are less correlated, implying two independent baseline features further improve PSM performance. Baseline features' incremental predictive value is the primary driver of PSM method effectiveness. The level of independence between these features has a minor impact on enhancing PSM effectiveness with incremental improvements. This observation is consistently demonstrated in another numerical experiment, which specifically focused on a continuous outcome resembling PFS (Figure S5E,F).

Besides tumor PD-L1 expression, tumor mutational burden has been identified as an orthogonal biomarker that is predictive of clinical benefits from immune checkpoint inhibitors.[36] Based on the above analysis, we speculate that combining these two relevant, but uncorrelated features will significantly improve PSM predictability.

TABLE 1 The features chosen through LASSO logistic regression and their correlation with the clinical outcome.

Feature	Correlation value	Feature weight from LASSO	*p*–value	Correlation type
PDL1	0.2085	0.3031	0.0003	CramersV
Lym	0.1731	0.2409	0.0025	PointBiserial
NEOIWTYP	0.1344	0.1789	0.0193	CramersV
K	0.1231	0.1025	0.0322	PointBiserial
PROT	−0.074	−0.1155	0.199	PointBiserial

Abbreviation: LASSO, least absolute shrinkage and selection operator.

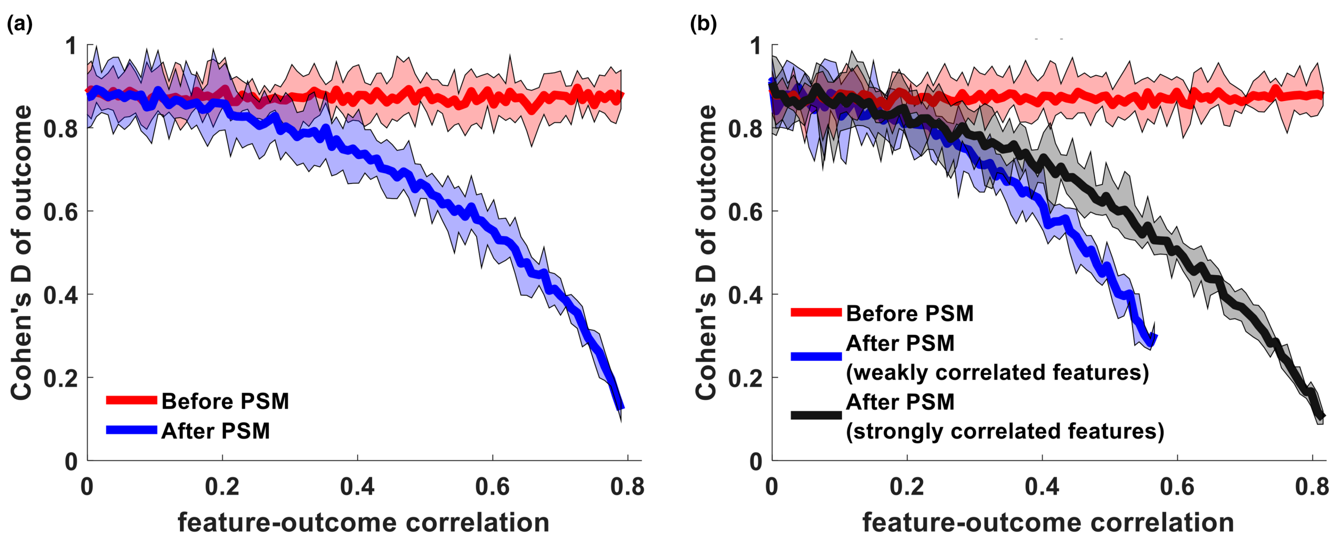

FIGURE 5 Effect of feature-outcome correlation on propensity score matching (PSM) performance when outcome is a binary variable. (a) Relation between feature-outcome correlation and Cohen's D of outcome when input is a single continuous feature. (b) Relation between feature-outcome correlation and Cohen's D of outcome when inputs are two features – one continuous variable and the other a binary variable. The filled region shows the complete range of values between the minimum and maximum of the simulations at each correlation and line shows the mean value of the simulations at each correlation.

Expanding the PSM database through synthetic data sharing

To make accurate predictions using PSM, it is essential to have ample historical data across various indications. However, sharing patient data among organizations is often difficult due to competition and privacy issues. A solution is to generate computationally synthesized datasets that mimic the structure and statistical properties of real historical data, including marginal distributions, feature-outcome correlations, covariance matrix, and feature correlations. The method has been extensively documented[37–41] and can successfully expand the external database, whereas ensuring the exclusion of any sensitive information.

How to put this PSM approach into practice?

Using our case study as an example, Gemstone-302 has established sugemalimab plus chemotherapy as a potential new standard first-line therapy for patients with stage IV NSCLC. This will serve as the historical control to which the results from a new single-arm phase IB trial involving an experimental therapy (e.g., PD-L1 bispecific antibody plus chemotherapy) is prospectively compared. As mentioned in case study 1, we have identified the key baseline features from Gemstone-302 (PDL1 and Lym) that will be used for PSM on the experimental arm using patient-level data in the phase IB study. Comparison of end points (e.g., ORR after PSM) provides a more informative assessment of the new treatment effect that may otherwise be confounded by an imbalance of predictive baseline features.

AUTHOR CONTRIBUTIONS

S.W.T., Y.C.S., R.P., M.C., and R.K. wrote the manuscript. S.W.T., R.K., and A.T. designed the research. S.W.T., Y.C.S., R.P., H.W., C.W., and M.C. performed the research. S.W.T., Y.C.S., R.P., and M.C. analyzed the data.

ACKNOWLEDGMENTS

The authors thank Bhairav Paleja for his insights on cancer pathophysiology that greatly informed the analysis. We express our gratitude to Melissa Curtis for her critical review.

FUNDING INFORMATION

This work was supported by CStone Pharmaceuticals.

CONFLICT OF INTEREST STATEMENT

S.-W.T., Y.-C.S., Y.S., H.W., C.W., and A.T. are employees of CStone Pharmaceuticals. All other authors declared no competing interests in this work.

ORCID

Shu-Wen Teng https://orcid.org/0009-0000-3102-8047
Ravikumar Pallantla https://orcid.org/0000-0002-4537-0991
Yucheng Sheng https://orcid.org/0000-0003-1460-5070

REFERENCES

1. Jemielita T, Tse A, Chen C. Oncology phase II proof-of-concept studies with multiple targets: randomized controlled trial or single arm? *Pharm Stat*. 2020;19:117-125.
2. Chen C. To go or not to go in phase 2/3 oncology trials, a critical question with a unified answer. *Contemp Clin Trials*. 2023;128:107146.
3. Sondak VK, Khushalani NI. Echoes of a failure: what lessons can we learn? *Lancet Oncol*. 2019;20:1037-1039.
4. O'Day SJ, Eggermont AMM, Chiarion-Sileni V, et al. Final results of phase III SYMMETRY study: randomized, double-blind trial of Elesclomol plus paclitaxel versus paclitaxel alone As treatment for chemotherapy-naive patients with advanced melanoma. *JCO*. 2013;31:1211-1218.
5. Mayawala K, de Alwis DP, Sachs JR. Model informed drug development: novel oncology agents are lost in translation. *Clin Cancer Res*. 2019;25:6564-6566.
6. Grayling MJ, Dimairo M, Mander AP, Jaki TF. A review of perspectives on the use of randomization in phase II oncology trials. *J Natl Cancer Inst*. 2019;111:1255-1262.
7. Thorlund K, Dron L, Park JJ, Mills EJ. Synthetic and external controls in clinical trials – a primer for researchers. *Clin Epidemiol*. 2020;12:457-467.
8. Hashmi M, Rassen J, Schneeweiss S. Single-arm oncology trials and the nature of external controls arms. *J Comp Eff Res*. 2021;10:1053-1066.
9. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar Behav Res*. 2011;46:399-424.
10. Fisher CK, Smith AM, Walsh JR, et al. Machine learning for comprehensive forecasting of Alzheimer's disease progression. *Sci Rep*. 2019;9:13622.
11. FDA guidance for industry: considerations for the design and conduct of externally controlled trials for drug and biological products. 2023. https://www.fda.gov/media/164960/download. Accessed July 6, 2023.
12. Kjeldsen JW, Lorentzen CL, Martinenaite E, et al. A phase 1/2 trial of an immune-modulatory vaccine against IDO/PD-L1 in combination with nivolumab in metastatic melanoma. *Nat Med*. 2021;27:2212-2223.
13. Zhou C, Wang Z, Sun Y, et al. Sugemalimab versus placebo, in combination with platinum-based chemotherapy, as first-line treatment of metastatic non-small-cell lung cancer (GEMSTONE-302): interim and final analyses of a double-blind, randomised, phase 3 clinical trial. *Lancet Oncol*. 2022;23:220-233.
14. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer*. 2009;45:228-247.
15. Andrillon A, Pirracchio R, Chevret S. Performance of propensity score matching to estimate causal effects in small samples. *Stat Methods Med Res*. 2020;29:644-658.
16. Bottigliengo D, Baldi I, Lanera C, et al. Oversampling and replacement strategies in propensity score matching: a critical

review focused on small sample size in clinical settings. *BMC Med Res Methodol*. 2021;21:256.
17. Claret L, Jin JY, Ferté C, et al. A model of overall survival predicts treatment outcomes with Atezolizumab versus chemotherapy in non–small cell lung cancer based on early tumor kinetics. *Clin Cancer Res*. 2018;24:3292-3298.
18. Stein WD, Gulley JL, Schlom J, et al. Tumor regression and growth rates determined in five intramural NCI prostate cancer trials: the growth rate constant as an indicator of therapeutic efficacy. *Clin Cancer Res*. 2011;17:907-917.
19. Gong J, Cao J, Zhang Q, et al. Safety, antitumor activity and biomarkers of sugemalimab in Chinese patients with advanced solid tumors or lymphomas: results from the first-in-human phase 1 trial. *Cancer Immunol Immunother*. 2022;71:1897-1908.
20. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163:1149-1156.
21. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41-55.
22. Xu Z, Kalbfleisch JD. Propensity score matching in randomized clinical trials. *Biometrics*. 2010;66:813-823.
23. Austin PC. Primer on statistical interpretation or methods report card on propensity-score matching in the cardiology literature from 2004 to 2006: a systematic review. *Circ Cardiovasc Qual Outcomes*. 2008;1:62-67.
24. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm Stat*. 2011;10:150-161.
25. Austin PC, Stuart EA. The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. *Stat Methods Med Res*. 2017;26:1654-1670.
26. Flury BK, Riedwyl H. Standard distance in univariate and multivariate analysis. *Am Stat*. 1986;40:249.
27. Hinshaw DC, Shevde LA. The tumor microenvironment innately modulates cancer progression. *Cancer Res*. 2019;79:4557-4566.
28. Cézé N, Thibault G, Goujon G, et al. Pre-treatment lymphopenia as a prognostic biomarker in colorectal cancer patients receiving chemotherapy. *Cancer Chemother Pharmacol*. 2011;68:1305-1313.
29. Huemer F, Lang D, Westphal T, et al. Baseline absolute lymphocyte count and ECOG performance score are associated with survival in advanced non-small cell lung cancer undergoing PD-1/PD-L1 blockade. *J Clin Med*. 2019;8:1014.
30. Lee YJ, Park YS, Lee HW, Park TY, Lee JK, Heo EY. Peripheral lymphocyte count as a surrogate marker of immune checkpoint inhibitor therapy outcomes in patients with non-small-cell lung cancer. *Sci Rep*. 2022;12:626.
31. Borghaei H, Paz-Ares L, Horn L, et al. Nivolumab versus docetaxel in advanced nonsquamous non–small-cell lung cancer. *N Engl J Med*. 2015;373:1627-1639.
32. Herbst RS, Baas P, Kim DW, et al. Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. *Lancet*. 2016;387:1540-1550.
33. Socinski MA, Jotte RM, Cappuzzo F, et al. Atezolizumab for first-line treatment of metastatic nonsquamous NSCLC. *N Engl J Med*. 2018;378:2288-2301.
34. Gandhi L, Rodríguez-Abreu D, Gadgeel S, et al. Pembrolizumab plus chemotherapy in metastatic non–small-cell lung cancer. *N Engl J Med*. 2018;378:2078-2092.
35. Höfler M. Causal inference based on counterfactuals. *BMC Med Res Methodol*. 2005;5:28.
36. Sha D, Jin Z, Budczies J, Kluck K, Stenzinger A, Sinicrope FA. Tumor mutational burden as a predictive biomarker in solid tumors. *Cancer Discov*. 2020;10:1808-1825.
37. Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. *BMC Med Res Methodol*. 2020;20:108.
38. Gambs S, Ladouceur F, Laurent A, Roy-Gaumond A. Growing synthetic data through differentially-private vine copulas. *Proc Priv Enh Technol*. 2021;2021:122-141.
39. Xie L, Lin K, Wang S, Wang F, Zhou J. Differentially private generative adversarial network. 2018 https://arxiv.org/abs/1802.06739.
40. Chen D, Cheung SS, Chuah C-N, Ozonoff S. Differentially private generative adversarial networks with model inversion. *In 2021 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE; 2021:1-6. doi:10.1109/WIFS53200.2021.9648378
41. Zhang J, Cormode G, Procopiuc CM, Srivastava D, Xiao X. PrivBayes: private data release via Bayesian networks. *ACM Trans Database Syst*. 2017;42:1-41.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Teng S-W, Su Y-C, Pallantla R, et al. Can a propensity score matching method be applied to assessing efficacy from single-arm proof-of-concept trials in oncology? *CPT Pharmacometrics Syst Pharmacol*. 2023;12:1347-1357. doi:10.1002/psp4.13014